The Bias-Variance Tradeoff: A Deep Dive

Nipun Batra and teaching staff

IIT Gandhinagar

August 15, 2025

Table of Contents

Understanding the Problem Setup

The Learning Problem: A Real-World Example

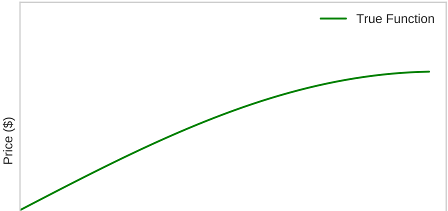
Definition: Our Scenario

Goal: Predict housing prices based on house area

Example: The True Relationship

Unknown to us: There exists a true function $f_{\theta_{\text{true}}}$ that perfectly relates area to price:

$$y_t = f_{\theta_{\text{true}}}(x_t)$$



The Three Sources of Prediction Error

Important: Fundamental Question

Why do our predictions fail? What causes the difference between our predictions and reality?

Definition: Three Universal Sources of Error

Every machine learning prediction suffers from:

1. **Noise** - Irreducible randomness in the data
2. **Bias** - Systematic errors from model assumptions
3. **Variance** - Sensitivity to particular training sets

Key Points

The Tradeoff: We can often reduce bias OR variance, but not both simultaneously!

Source 1: Noise - The Irreducible Error

Understanding Noise: The Fundamental Limitation

Definition: What is Noise?

Noise represents factors affecting the target that we cannot observe or control

Example: Real-World Noise Sources

In housing prices:

- House condition (hard to measure precisely)
- Neighborhood market dynamics
- Buyer's personal preferences

Noise: Why It's Irreducible

Example: More Noise Sources

Additional factors we cannot control:

- Economic conditions on sale day
- Unmeasurable aesthetic factors
- Random market fluctuations
- Measurement errors in data collection

Important: Key Insight

Irreducible Error: No matter how sophisticated our model, noise cannot be eliminated!

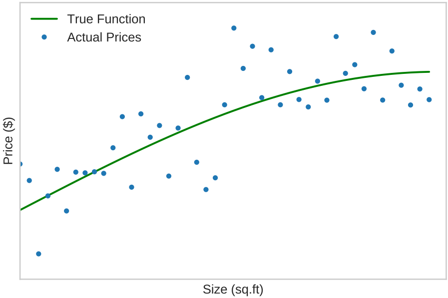
Noise: Mathematical Formulation

Key Points

The Noisy Relationship **True relationship becomes:**

$$y_t = f_{\theta_{\text{true}}}(x_t) + \epsilon_t$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is the noise term



Noise: Mathematical Properties

Definition: Key Properties of Noise

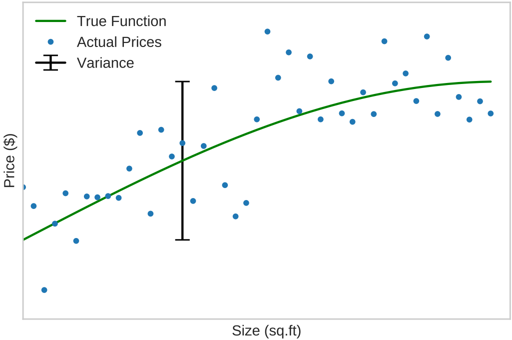
- **Zero mean:** $E[\epsilon_t] = 0$ (unbiased)
- **Constant variance:** $\mathrm{Var}(\epsilon_t) = \sigma^2$
- **Independent:** Each observation's noise is independent

Key Points

Why These Properties Matter

- **Zero mean:** Noise doesn't systematically bias our target
- **Constant variance:** Prediction uncertainty is consistent
- **Independence:** One data point's noise doesn't affect others

Visualizing Noise: Data Distribution



Key Points

Key Observation:

- Data points scatter around the true function
- The spread (variance) is constant: σ^2

Source 2: Bias - Systematic Model Limitations

Understanding Bias: Model Flexibility

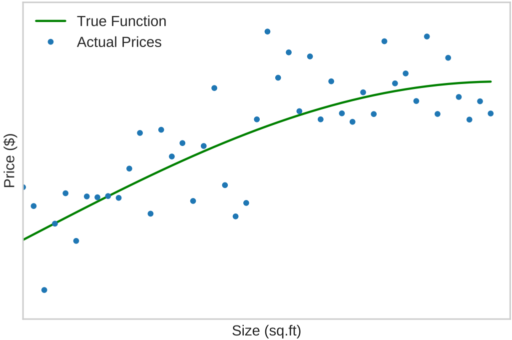
Definition: What is Bias?

Bias measures how well our model class can represent the true function

Example: Extreme Example: Constant Function

Model choice: $\hat{f}(x) = c$ (constant, regardless of house size)
Question: Can this model capture the true price-size relationship?

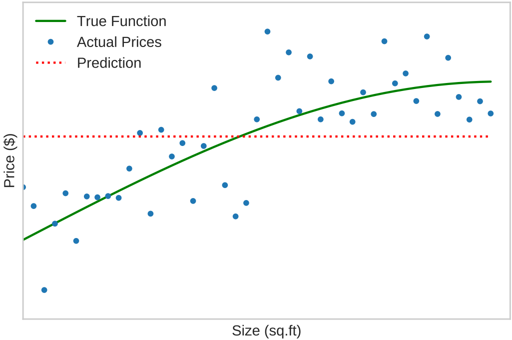
Bias: Visualizing the Problem



Important:

Obvious Problem: A constant function cannot capture any relationship with house size!

Bias: Fitting a Constant Model

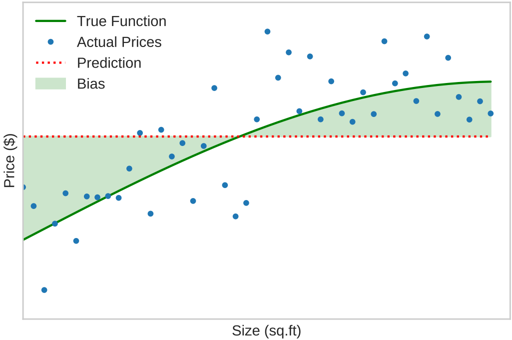


Key Points

Best Constant Fit:

- The optimal constant is the average of all prices
- But this completely ignores the size information!

Bias: Visualizing the Systematic Error



Definition: Bias Definition

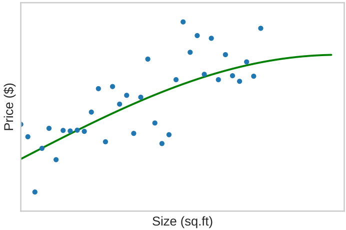
$$\text{Bias}(x) = f_{\theta_{\text{true}}}(x) - E[\hat{f}(x)]$$

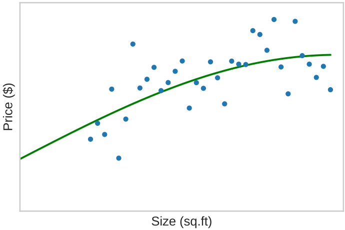
The systematic difference between truth and average prediction

Multiple Datasets: Understanding Variability

Key Points

Crucial Insight: Many different datasets are possible from the same true relationship!



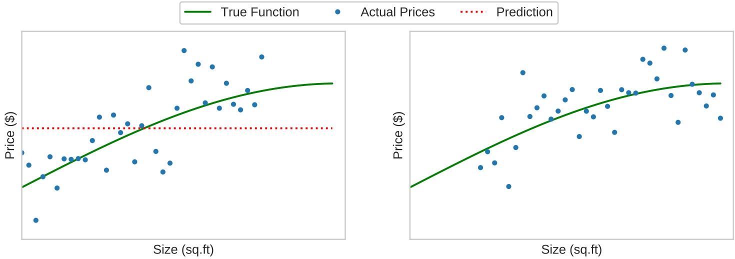


Example:

Same underlying relationship, different data points due to:

- Random sampling of houses

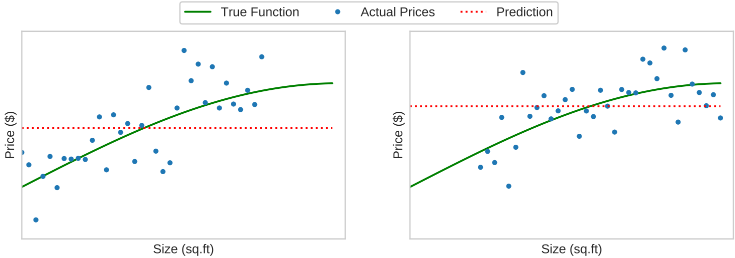
Fitting Models to Different Datasets



Key Points

Question: If we fit the same model type (constant) to different datasets, what happens?

Different Predictions from Different Datasets



Important:

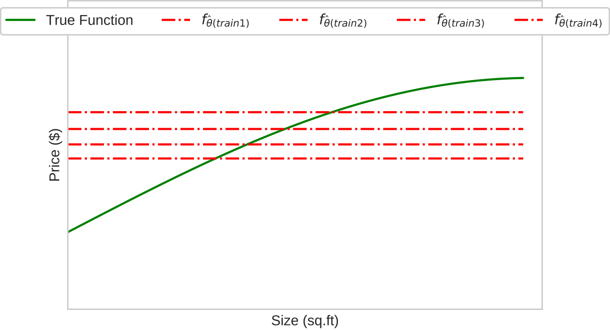
Key Observation: Even with the same model type, we get different predictions!

Definition:

This variability leads us to two concepts:

- **Average prediction:** What happens "on average" across all

Many Datasets: The Full Picture



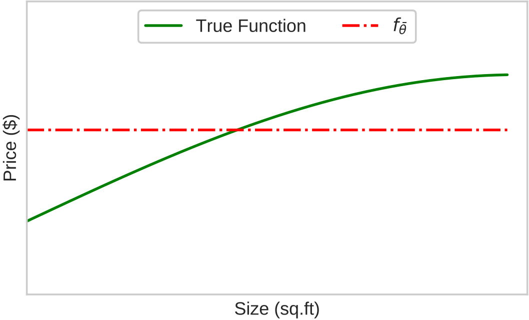
Key Points

Multiple Datasets: Each gives a slightly different constant fit

Example:

The Big Question: What is the "typical" or "expected" prediction

The Average Model: Expected Prediction



Definition: Expected Prediction

$$E[\hat{f}(x)] = \text{Average prediction across all possible training sets}$$

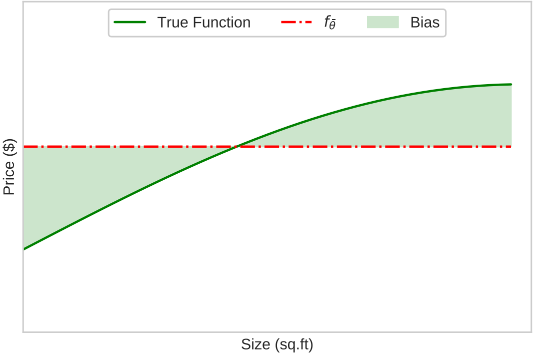
Key Points

Bias: The Final Definition

Definition: Bias Formula

$$\text{Bias}(x) = f_{\theta_{\text{true}}}(x) - E[\hat{f}(x)]$$

Difference between truth and expected prediction

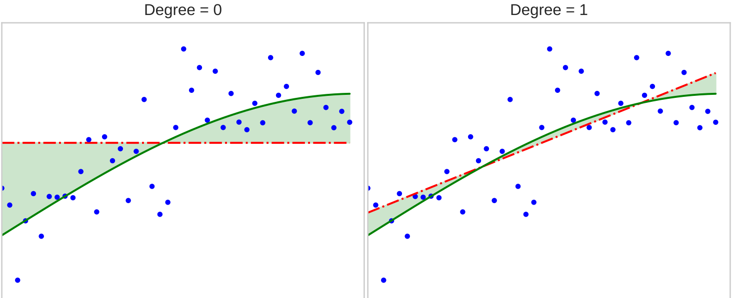


Model Complexity vs Bias: The Relationship

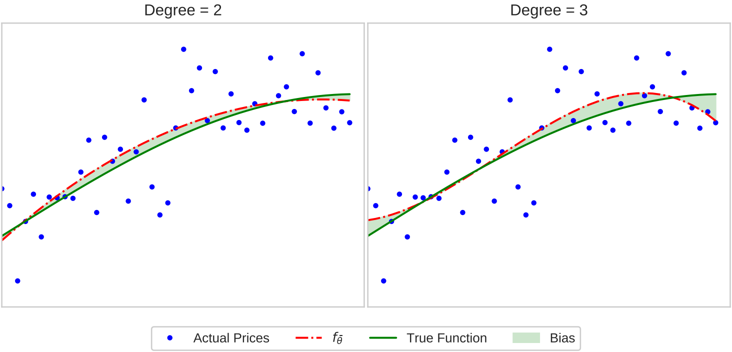
Key Points

Universal Pattern:

- **Increase complexity** $\rightarrow$ Model becomes more flexible
- **More flexible** $\rightarrow$ Can better approximate true function
- **Better approximation** $\rightarrow$ Bias decreases



High-Complexity Models: Near-Zero Bias



Important:

High-Degree Polynomials: Can approximate almost any smooth function!

Definition:

Source 3: Variance - Dataset Sensitivity

From Bias to Variance: The Other Side

Important:

We've seen: High-complexity models have low bias
Question: If low bias is good, why not always use high-complexity models?

Definition: Enter Variance

Variance measures how much predictions change when we train on different datasets

Key Points

Intuition:

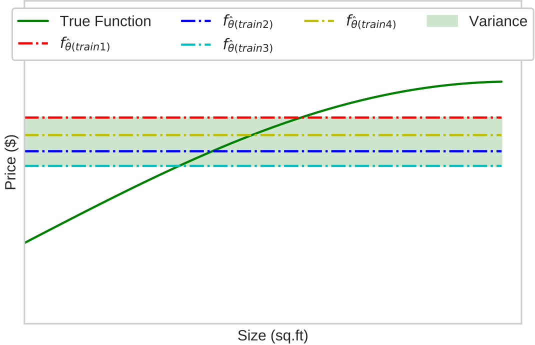
- Simple models: Stable, consistent predictions
- Complex models: Highly sensitive to specific training data

Understanding Variance: Prediction Consistency

Definition: Variance Definition

Variance $=$ How much do predictions vary across different training sets?

$$\text{Var}(\hat{f}(x)) = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$

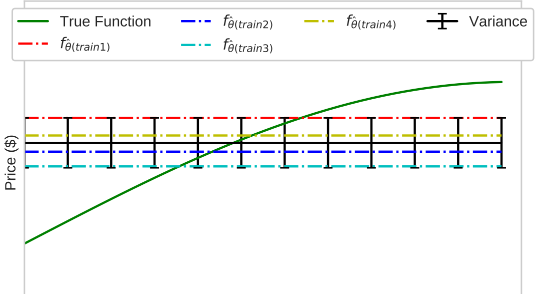


Low Complexity: Low Variance

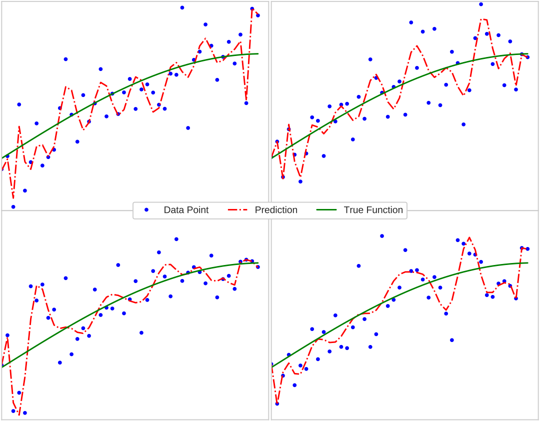
Key Points

Simple Models (e.g., linear):

- Few parameters to estimate
- Robust to data variations
- Consistent predictions



High Complexity: The Variance Problem Emerges



Important:

Warning Signs:

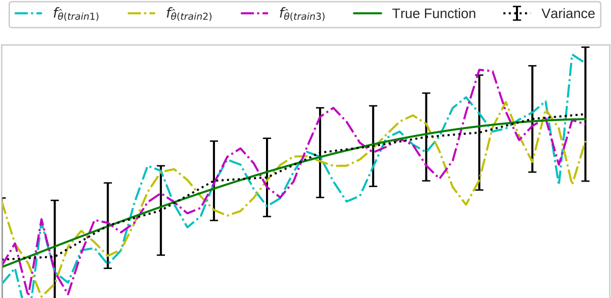
- Models look very different across datasets

High Complexity: Extreme Variance

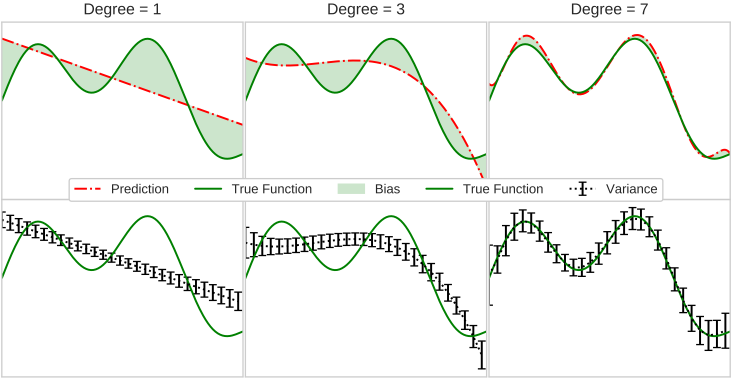
Key Points

Complex Models (e.g., high-degree polynomials):

- Many parameters to estimate
- Overfit to specific training data
- Dramatically different predictions



The Bias-Variance Tradeoff: The Central Tension



Important: The Fundamental Tradeoff

- **Simple models:** High bias, low variance
- **Complex models:** Low bias, high variance
- **Optimal complexity:** Balance between the two

Mathematical Derivation: The Bias-Variance Decomposition

Why Mathematical Analysis Matters

Definition: The Goal

Question: Can we mathematically prove that error = bias^2 + variance + noise?

Key Points

Why This Matters

- **Theoretical foundation:** Understand the fundamental nature of learning
- **Model selection:** Know exactly what we're trading off
- **Algorithm design:** Create methods that explicitly balance bias and variance

Example: Expected Error Across All Possible Datasets

Setting Up the Mathematical Framework

Definition: What We Want to Prove

$$E[\text{Error}] = \text{Noise} + \text{Bias}^2 + \text{Variance}$$

Key Points

Our Approach

1. Start with prediction error at a single point
2. Use squared loss: $(y - \hat{f}(x))^2$
3. Take expectation over all sources of randomness
4. Apply algebraic manipulation to separate terms

Example: Sources of Randomness

- **Training set:** Which data points we observe

Mathematical Setup: Defining the Components

Definition: True Relationship with Noise

$$y = f_{\text{true}}(x) + \epsilon \quad \text{where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Key Points

Key Definitions

- $f_{\text{true}}(x)$: The unknown true function
- $\hat{f}(x)$: Our model's prediction (depends on training data)
- $E[\hat{f}(x)]$: Expected prediction over all possible training sets
- ϵ: Irreducible noise with variance σ^2

Formal Definitions: Bias and Variance

Definition: Bias

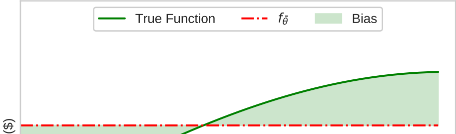
$$\text{Bias}(x) = f_{\text{true}}(x) - E[\hat{f}(x)]$$

Systematic error: Difference between truth and expected prediction

Definition: Variance

$$\text{Variance}(x) = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$

Prediction instability: Expected squared deviation from mean prediction



The Main Theorem: Bias-Variance Decomposition

Important: The Fundamental Result

For any point x and any learning algorithm:

$$E[(y - \hat{f}(x))^2] = \sigma^2 + [\text{Bias}(x)]^2 + \text{Variance}(x)$$

Definition: Component Interpretation

- σ^2: **Irreducible error** (noise)
- $[\text{Bias}(x)]^2$: **Systematic error** (underfitting)
- $\text{Variance}(x)$: **Random error** (overfitting)

Key Points

Coming Up: We'll prove this step-by-step using careful algebraic manipulation

Starting the Proof: Expected Squared Error

Definition: What We're Proving

$$E[(y - \hat{f}(x))^2] = \sigma^2 + [\text{Bias}(x)]^2 + \text{Variance}(x)$$

Key Points

Our Strategy

1. Start with squared error: $(y - \hat{f}(x))^2$
2. Add and subtract strategic terms
3. Expand and use linearity of expectation
4. Show cross-terms cancel out
5. Identify the three components

Example: Key Insight

Step 1: Setting Up the Expectation

Definition: Squared Loss at Point x

Individual prediction error: $(y - \hat{f}(x))^2$

Key Points

Taking Expectations **Expected error over all randomness:**

$$E_{\mathcal{D},y}[(y - \hat{f}(x))^2]$$

where:

- $\mathcal{D}$: Random training set
- y: Random target (includes noise)

Example: Why Two Sources of Randomness?

Step 2: The Add-and-Subtract Trick

Key Points

Starting Point

$$E[(y - \hat{f}(x))^2]$$

Example: Strategic Addition and Subtraction

Add and subtract $f_{\text{true}}(x)$:

$$E[(y - f_{\text{true}}(x) + f_{\text{true}}(x) - \hat{f}(x))^2]$$

Definition: Grouping Terms

$$E[\underbrace{(y - f_{\text{true}}(x))}_{\text{noise: } \epsilon} + \underbrace{(f_{\text{true}}(x) - \hat{f}(x))}_{\text{prediction error}}]^2$$

Important:

Step 3A: Expanding the Square

Key Points

Starting with Our Expression

$$E[\underbrace{(y - f_{\text{true}}(x))}_{\epsilon} + \underbrace{(f_{\text{true}}(x) - \hat{f}(x))}_{\text{prediction error}}]^2$$

Example: Apply $(a+b)^2 = a^2 + 2ab + b^2$

Let: $a = \epsilon$ and $b = (f_{\text{true}}(x) - \hat{f}(x))$
Then: $(a+b)^2 = a^2 + 2ab + b^2$

Step 3B: The Expanded Form

Definition: After Expansion

$$E[\epsilon^2 + 2\epsilon(f_{\text{true}}(x) - \hat{f}(x)) + (f_{\text{true}}(x) - \hat{f}(x))^2]$$

Key Points

Apply Linearity of Expectation

$$E[\epsilon^2] + 2E[\epsilon(f_{\text{true}}(x) - \hat{f}(x))] + E[(f_{\text{true}}(x) - \hat{f}(x))^2]$$

Step 3C: Naming the Three Terms

Definition: Our Three Terms

- **Term 1:** $E[\epsilon^2]$ (the noise term)
- **Term 2:** $2E[\epsilon(f_{\text{true}}(x) - \hat{f}(x))]$ (cross-term)
- **Term 3:** $E[(f_{\text{true}}(x) - \hat{f}(x))^2]$ (prediction error)

Key Points

Our Strategy **We'll analyze each term separately:**

- Term 1 $\rightarrow$ Noise (σ^2)
- Term 2 $\rightarrow$ Will be zero!
- Term 3 $\rightarrow$ Bias^2 + Variance

Step 4A: Analyzing Term 1 - Setup

Definition: Term 1 Recall

$$\text{Term } 1 = E[\epsilon^2]$$

where $\epsilon = y - f_{\text{true}}(x)$ is the noise

Key Points

Key Insight **Independence:** The noise ϵ doesn't depend on our training set!

- Noise is a property of the data generation process
- Training set selection doesn't affect noise level
- Noise is the same regardless of which model we choose

Step 4B: Term 1 - The Calculation

Example: By Definition of Noise

$$\epsilon = y - f_{\text{true}}(x) \sim \mathcal{N}(0, \sigma^2)$$

Key Points

Therefore

$$E[\epsilon^2] = \text{Var}(\epsilon) + (E[\epsilon])^2 = \sigma^2 + 0^2 = \sigma^2$$

Important: Result

$$\boxed{\text{Term } 1 = \sigma^2}$$

This is our irreducible error (noise)!

Step 5A: Analyzing Term 2 - Setup

Definition: Term 2 Recall

$$\text{Term 2} = 2E[\epsilon(f_{\text{true}}(x) - \hat{f}(x))]$$

Key Points

Key Independence Property **Crucial insight:** ϵ (noise) is independent of $\hat{f}(x)$ (our prediction)

- Noise occurs in nature, regardless of our model
- Our model $\hat{f}$ depends only on training data
- Training data and future noise are independent

Step 5B: Term 2 - Why Independence Matters

Example: What Independence Means

If X and Y are independent: $E[XY] = E[X] \cdot E[Y]$
In our case:

- $X = \epsilon$ (noise)
- $Y = f_{\text{true}}(x) - \hat{f}(x)$ (prediction error)

Key Points

Apply Independence

$$E[\epsilon(f_{\text{true}}(x) - \hat{f}(x))] = E[\epsilon] \cdot E[f_{\text{true}}(x) - \hat{f}(x)]$$

Step 5C: Term 2 - The Final Calculation

Example: Using $E[\epsilon] = 0$

$$E[\epsilon(f_{\text{true}}(x) - \hat{f}(x))] = E[\epsilon] \cdot E[f_{\text{true}}(x) - \hat{f}(x)]$$
$$= 0 \cdot E[f_{\text{true}}(x) - \hat{f}(x)] = 0$$

Important: Result

$$\boxed{\text{Term } 2 = 2 \times 0 = 0}$$

The cross-term vanishes completely!

Key Points

Why This Matters Cross-terms often make math messy, but here they cancel out beautifully!

Step 6: Analyzing Term 3 - The Prediction Error

Definition: Term 3 Analysis

$$E[(f_{\text{true}}(x) - \hat{f}(x))^2]$$

Key Points

Another Independence **Key insight:** $(f_{\text{true}}(x) - \hat{f}(x))$ doesn't depend on the noise ϵ

- $f_{\text{true}}(x)$ is deterministic
- $\hat{f}(x)$ depends only on training inputs/outputs (not future noise)

Example: Simplification

$$E[(f_{\text{true}}(x) - \hat{f}(x))^2] = \text{MSE of prediction}$$

Interim Summary: Progress So Far

Definition: What We Have

$$E[(y - \hat{f}(x))^2] = \sigma^2 + 0 + E[(f_{\text{true}}(x) - \hat{f}(x))^2]$$
$$= \sigma^2 + E[(f_{\text{true}}(x) - \hat{f}(x))^2]$$

Key Points

Next Challenge **Goal:** Decompose $E[(f_{\text{true}}(x) - \hat{f}(x))^2]$ into bias^2 + variance

Example: Strategy for Next Step

Another add-and-subtract trick: We'll add and subtract $E[\hat{f}(x)]$ inside the MSE term

Step 7A: The Second Decomposition - Setup

Key Points

Current Status

$$E[(y - \hat{f}(x))^2] = \sigma^2 + E[(f_{\text{true}}(x) - \hat{f}(x))^2]$$

Example: Our Next Challenge

Goal: Break down $E[(f_{\text{true}}(x) - \hat{f}(x))^2]$ into bias^2 + variance

Key Points

Strategy **Another add-and-subtract trick!** We'll use $E[\hat{f}(x)]$ (the expected prediction)

Step 7B: The Second Add-and-Subtract Trick

Definition: Starting Point

$$E[(f_{\text{true}}(x) - \hat{f}(x))^2]$$

Example: Add and Subtract $E[\hat{f}(x)]$

$$E[(f_{\text{true}}(x) - E[\hat{f}(x)] + E[\hat{f}(x)] - \hat{f}(x))^2]$$

Key Points

Grouping the Terms

$$E[\underbrace{(f_{\text{true}}(x) - E[\hat{f}(x)])}_{\text{bias}} + \underbrace{(E[\hat{f}(x)] - \hat{f}(x))}_{\text{variance deviation}}]^2$$

Important:

Step 8A: Setting Up the Final Expansion

Key Points

Our Current Expression

$$E[(\text{bias} + \text{variance deviation})^2]$$

Definition: Let's Define Clearly

- $\alpha = f_{\text{true}}(x) - E[\hat{f}(x)]$ (the bias)
- $\beta = E[\hat{f}(x)] - \hat{f}(x)$ (deviation from expected prediction)

Key Points

What We're About to Do Expand $(\alpha + \beta)^2$ and analyze each term separately

Step 8B: Expanding the Square

Example: Using $(a+b)^2 = a^2 + 2ab + b^2$

$$E[(\alpha+\beta)^2] = E[\alpha^2 + 2\alpha\beta + \beta^2]$$

Key Points

Apply Linearity of Expectation

$$E[\alpha^2] + 2E[\alpha\beta] + E[\beta^2]$$

Definition: Three Terms to Analyze

- **Term A:** $E[\alpha^2]$
- **Term B:** $2E[\alpha\beta]$
- **Term C:** $E[\beta^2]$

Step 9A: Analyzing Term A - The Bias Term

Definition: Term A Recall

$$E[\alpha^2] = E[(f_{\text{true}}(x) - E[\hat{f}(x)])^2]$$

where $\alpha = f_{\text{true}}(x) - E[\hat{f}(x)]$

Key Points

Critical Insight α **is deterministic (not random)!**

- $f_{\text{true}}(x)$ is a fixed function value
- $E[\hat{f}(x)]$ is the expected prediction (a constant)

Step 9B: Why Deterministic Matters

Example: When Something is Deterministic

If c is a constant: $E[c] = c$
In our case: $\alpha = f_{\text{true}}(x) - E[\hat{f}(x)]$ is constant

Key Points

Therefore

$$E[\alpha^2] = E[(f_{\text{true}}(x) - E[\hat{f}(x)])^2] = (f_{\text{true}}(x) - E[\hat{f}(x)])^2$$

Important: Result

$$\boxed{E[\alpha^2] = [\text{Bias}(x)]^2}$$

First component: Bias squared!

Step 10A: Analyzing Term B - The Cross-Term

Definition: Term B Recall

$$E[\alpha\beta] = E[(f_{\text{true}}(x) - E[\hat{f}(x)])(E[\hat{f}(x)] - \hat{f}(x))]$$

Key Points

Key Insight α **is deterministic:** $(f_{\text{true}}(x) - E[\hat{f}(x)])$ is a constant

- Can factor constants out of expectations
- $E[c \cdot X] = c \cdot E[X]$ when c is constant

Step 10B: Factoring Out the Constant

Example: Using the Constant Rule

$$E[\alpha\beta] = E[(f_{\text{true}}(x) - E[\hat{f}(x)])(E[\hat{f}(x)] - \hat{f}(x))]$$
$$= (f_{\text{true}}(x) - E[\hat{f}(x)]) \cdot E[E[\hat{f}(x)] - \hat{f}(x)]$$

Key Points

Simplifying the Expectation

$$E[E[\hat{f}(x)] - \hat{f}(x)] = E[\hat{f}(x)] - E[\hat{f}(x)] = 0$$

Important: Result

$$\boxed{E[\alpha\beta] = \text{bias} \times 0 = 0}$$

Cross-term vanishes again!

Step 11A: Analyzing Term C - The Variance Term

Definition: Term C Recall

$$E[\beta^2] = E[(E[\hat{f}(x)] - \hat{f}(x))^2]$$

where $\beta = E[\hat{f}(x)] - \hat{f}(x)$

Key Points

Does This Look Familiar? **Compare with the definition of variance:**

$$\text{Variance}(X) = E[(X - E[X])^2]$$

Step 11B: Recognizing the Variance Formula

Example: Rewriting Term C

$$E[\beta^2] = E[(E[\hat{f}(x)] - \hat{f}(x))^2]$$
$$= E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$

Key Points

This is Exactly...

$$\text{Variance}(\hat{f}(x)) = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$

Important: Result

$$\boxed{E[\beta^2] = \text{Variance}(\hat{f}(x))}$$

Final component: Variance!

The Complete Bias-Variance Decomposition

Important: Putting It All Together

$$E[(y - \hat{f}(x))^2] = \sigma^2 + [\text{Bias}(x)]^2 + \text{Variance}(x)$$

Definition: Component Summary

- σ^2 = **Irreducible error** (noise in data)
- $[\text{Bias}(x)]^2$ = **Systematic error** (model assumptions)
- $\text{Variance}(x)$ = **Random error** (training set sensitivity)

Key Points

The Fundamental Tradeoff

- **Reduce bias:** Use more complex models → Increase variance
- **Reduce variance:** Use simpler models → Increase bias

Summary: The Bias-Variance Tradeoff

Definition: What We've Proven

Every prediction error can be decomposed as:

$$\text{Total Error} = \text{Noise} + \text{Bias}^2 + \text{Variance}$$

Key Points

Key Takeaways

- **Noise:** Cannot be reduced (irreducible)
- **Bias:** Reduced by increasing model complexity
- **Variance:** Reduced by decreasing model complexity
- **Optimal model:** Balances bias and variance

Important: Practical Applications